

UNCLASSIFIED

AD NUMBER
AD866777
NEW LIMITATION CHANGE
TO Approved for public release, distribution unlimited
FROM Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; JAN 1970. Other requests shall be referred to Commanding Officer, Edgewood Arsenal, Attn: SMUEA-TSTI-T, Edgewood Arsenal, MD 21010.
AUTHORITY
USAEA Notice, 8 May 1970

THIS PAGE IS UNCLASSIFIED

AD 866777

AD

A CHEMICAL INFORMATION AND DATA SYSTEM

Semiannual Report

by

Clarence T. Van Meter

Ruth V. Powers

Helen N. Hill

Margaret Milne

Nancy Hamp

T. C. Chen

Morris Plotkin

31 January 1970



DEPARTMENT OF THE ARMY

EDGEWOOD ARSENAL

Technical Support Directorate

Technical Data & Value Engineering Management Office

Edgewood Arsenal, Maryland 21010

Contract DAAA15-69-C-0140

UNIVERSITY OF PENNSYLVANIA

PHILADELPHIA PENNSYLVANIA 19104

480255-01	
REF	WHITE SECTION <input type="checkbox"/>
SEC	WHITE SECTION <input checked="" type="checkbox"/>
CLASSIFICATION	
DATE	
BY	
REMARKS	
2	

Distribution Statement

This document is subject to special export controls and each transmittal to a foreign government or a foreign national may be made only with prior approval of the Commanding Officer, Edgewood Arsenal, ATTN: SMUEA-TSTI-T, Edgewood Arsenal, Maryland 21010.

Disclaimer

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Disposition

Destroy this report when no longer needed. Do not return it to the originator.

A CHEMICAL INFORMATION AND DATA SYSTEM

Semiannual Report

by

Clarence T. Van Meter, Ruth V. Powers,

Helen N. Hill, Margaret Milne,

Nancy Hamp, T. C. Chen, and Morris Plotkin

31 January 1970

Distribution Statement

This document is subject to special export controls and each transmittal to a foreign government or a foreign national may be made only by prior approval of the Commanding Officer, Edgewood Arsenal, ATTN: SMUEA-TSTI-T, Edgewood Arsenal, Maryland 21010

DEPARTMENT OF THE ARMY
EDGEWOOD ARSENAL

Technical Support Directorate
Technical Data & Value Engineering Management Office
Edgewood Arsenal, Maryland 21010

Contract DAAA15-69-C-0140

Task 2P062101A72702

UNIVERSITY OF PENNSYLVANIA
Philadelphia, Pennsylvania 19104

FOREWORD

The work described in this report was authorized under Task 2P062101A72702, Army Chemical Information and Data Systems (U). The work was started in July 1964 and is continuing. The information contained in this report represents work accomplished during the period 1 September 1969 through 31 January 1970.

The information in this document has not been cleared for release to the general public.

Acknowledgment

The project is pleased to acknowledge the generous cooperation of members of the staff of the Technical Data & Value Engineering Management Office and the EDP Systems personnel, Edgewood Arsenal in the conduct of various phases of the work.

Reproduction Statement

Reproduction of this document in whole or in part is prohibited except with permission of CO, Edgewood Arsenal, ATTN: SMUEA-TSTD, Edgewood Arsenal, Maryland 21010; however, Defense Documentation Center is authorized to reproduce the document for US Government purposes.

DIGEST

This document describes the research and development activities conducted on Project CIDS of the University of Pennsylvania during the period 1 September 1969 - 31 January 1970. All of the activities are designed toward either completion or advancement of the initial version of a model operational system.

File building during the period has been confined to the Edgewood Arsenal Task 07 file because these compounds are being tagged with both structural and nonstructural search keys and are thus searchable on either or both of these descriptive categories. To date, 13,985 candidate compounds for the file have been received; 10,666 of these have been registered. The average registered compound was tagged with four Edgewood Arsenal nonstructural descriptors.

All computer programs needed to assign the new (CIDS No. 6) chemical search screens have been written and tested, and the screens have been assigned successfully to a test file of 7226 Task 07 compounds. The average compound received 31 screen assignments in contrast to 23 using the experimental (CIDS No. 4) screens. The earlier ring finding program has been augmented with another program and the combination enables finding the rings in any smallest set present in a multiring system rather than finding only the rings in a prescribed smallest set such as those used in The Ring Index. The total lexicon of CIDS No. 6 screens will soon be assigned to the total registered Task 07 file and also to the 33,000 TOXINFO and CBCC compound file.

The improved atom-by-atom search program has been incorporated into the search system and, as a result of this and related modifications, the entire search system can now be accommodated in three phases rather than four. Other improvements in the search system include (a) the use of parenthesized logical expressions, (b) the incorporation of useful changes in the teletype command language, and (c) the ability to search on the basis of the six-digit number assigned to each compound at registration. The disk space required for a compound has been materially reduced by application of a more compact record format.

The program organization of the cathode ray tube terminal system is being basically altered to correct erratic performance traceable to problems in timing and transfer of control. An operating system has been designed for the DEC-338 programmed buffered display, suitable for the display of chemical compounds and

fragments, as used in CIDS real time search. Various program segments operating under a graphics monitor have been written to allow drawing and manipulating molecules on the screen, and a program has also been written for processing queries typed in at the DEC-338 teletype.

A document has been produced which describes and illustrates the rules for encoding queries addressed to the model operational CIDS. It also (a) discusses certain features of system operation and use to assist in the formulation of maximally efficient queries, (b) displays a standard query coding form, and (c) provides completed coding forms for 20 sample user-type questions.

Chemical specifications for the CIDS treatment of the following classes of compounds have been documented: (a) polypeptides, (b) polymers (other than natural types), and (c) metal-containing organics. In arriving at the specifications, due consideration was given to compatibility with (a) practices recommended by the international community of chemists, and (b) existing and planned CIDS computer programs. In addition to setting forth the specifications, the documents identify the search parameters and provide numerous illustrations of the chemical contents of compound records in CIDS format.

Additional documentation accomplished during the period include (a) detailed instructions for CIDS usage of the DURA chemical typewriter, and (b) definition at the programmer level of the CIDS master files. The latter includes the description of ten CIDS files. Sample tape dumps accompany the descriptions. Interim documentation of the overall system, in detail sufficient to disclose the types and magnitude of effort required to complete the system and convert it to the sponsor's computer, is currently in process.

TABLE OF CONTENTS

	page
1. Introduction	7
2. File Generation Techniques	7
3. The Search System	9
4. Status of CIDS Compound File	12
5. Identification of Rings in Cyclic Systems	15
6. New Screen Assignment	17
7. Cathode Ray Tube Research and Development	19
8. Documentation: Query Formulation and Encoding	20
9. Documentation: ACT III Chemical Typing Conventions	21
10. Documentation: Definition of CIDS Master Files	21
11. Documentation: Proposed CIDS Chemical Specifications for Polypeptides	22
12. Documentation: The CIDS Treatment of Synthetic Polymeric Substances	24
13. Documentation: The CIDS Treatment of (1) Metal-Containing Organic Compounds and (2) Esters	25
14. CIDS Documentation Scheduling Requirements	26
Literature Cited	27
Distribution List	29
Document Control Data - R&D, DD Form 1473, With Abstract and Keyword List	33

A CHEMICAL INFORMATION AND DATA SYSTEM
Semiannual Report

1. Introduction

All of the tasks undertaken during this contract period contribute, in one way or another, to either (a) completion of the initial version of a model operational system suitable for trial and demonstration purposes, or (b) evolution of the initial version of the system to a more advanced state. In addition to the usual activities associated with continued file building, these tasks involve (a) effecting improvements in file generation techniques and the search system, (b) replacing the experimental (CIDS No. 4) chemical search screens by the improved (CIDS No. 6) screens, (c) continuing R&D toward implementing the cathode ray tube as a component of the system, (d) documenting the required CIDS chemical specifications for types of compounds currently inadmissible to the file, and (e) providing documentation on the overall system sufficient for planning for conversion to the sponsor's Scientific and Engineering (S&E) computer.

The testing and processing associated with some of the above activities, together with the required preplanning and the assay of the findings, account for a large fraction of the Project's total effort. It would serve no useful purpose to provide a detailed account of these day-to-day operations; rather, the ultimate net results find expression only in terms of reported improvements in strategies, programs, techniques, etc.

2. File Generation Techniques

Advances in file generation techniques subdivide into the three areas indicated below.

Registration - A subroutine has been added to the REGISTRY system enabling it to take maximum advantage of the basic logic of the new atom-by-atom (A/A) search. This system employs A/A search to identify compounds submitted for registration, which have the same structural formula as a compound already on file. For each such match, a chemist determines from auxiliary data which of the submitted records are duplicates, which are updates and which are stereoisomers of the registered compound. The new subroutine renumbers the atoms of the structural formulas so that matching structural formulas are singled out most rapidly.

PRECEDING PAGE BLANK

Assignment of New Screens - Of the 10,666 compounds registered thus far in the Task 07 file, 7226 have been screened against the revised (CIDS No. 6) lexicon of molecular formula and structural fragment keys. The frequency of occurrence of various chemical features implied by the relative number of assignments of these keys is as expected for any sizable unbiased organic file. A more detailed analysis of the results of this screening is provided in Section 6 of this report.

One troublesome area of key assignment was the establishment of automatable criteria for identifying all rings in a cyclic system which should be assigned. The definition that was arrived at and subsequently incorporated into the ring screening programs is described in Sec. 5.

The currently expanding Task 07 file of compounds differs from previous CIDS files in that keys are employed not only for features of structure and molecular formula, but also for certain nonstructural information. A compound record which responds to one of the nonstructural information keys identifies one or more reference documents from which this information can be obtained. In Sec. 4 of this report, the set of Edgewood Arsenal nonstructural keys currently employed in this experimental approach to nonstructural data is listed and discussed in further detail.

The ability to search on the basis of the six-digit number assigned to each compound at registration has been implemented in an interim fashion. This capability might have been accomplished by assigning distinct registry number (RN) keys to each compound. The key index, however, is limited by core size as to the total number of different keys that can be accommodated, and could not accept a different key for each compound in the file. A solution was found to be the assignment of two keys to each compound, one designating the three high-order digits of the registry number and the other designating the three low-order digits. Thus, each pair of one high- with one low-order key corresponds to exactly one compound. Plans have been made to enable query specification of the complete number in a single key which will be expanded into two keys internally.

The high-order digits RN keys serve as a means of separating the file into blocks of 1000 compounds. This is a useful capability because the complete search system is designed to handle only the first 1000 responses to the keys. A query obtaining more than 1000 such responses can now be broken down into several

queries each of which uses the same original set of keys plus a high-order RN key. Each of these subqueries is thereby restricted to a different block of 1000 compounds and a manageable number of responses in each case is insured.

Revised File Storage Technique - Conversion of the 33,270 compound TOXINFO-CBCC file to the revised record format used in the Task 07 file has demonstrated the appreciable storage-saving capability of the new format as compared with the former method. In the new format, (a) the structural formula is stored in MCC (mechanical chemical code), a linear notation for encoding two dimensional chemical structures; (b) the characters in the SFI (structural formula image), which preserves the structural formula for output, are packed; and (c) the keys assigned to a compound are no longer stored as part of each record.

As a result of these innovations, the disk space required for the 33K compounds search file using the old (CIDS No. 4) screens was reduced to 40% of the amount required under the old format. These changes did not, of course, have any effect on the space requirements for the index of keys for this file. Based on the present disk space allotted for the file as compared with that allotted for the index, the index has now become the limiting factor in determining the total number of compounds which can be stored for real time search and will be somewhat more limiting using the new (CIDS No. 6) screens.

3. The Search System

Effort toward expanding and refining the search system subdivides into the three areas indicated below.

Incorporation of the New Atom-by-Atom Search Program - The new atom-by-atom (A/A) search and the programs which blow up the MCC to the FORTRAN connection table (CT) have been incorporated into the search system and the integrated system is now being debugged. Various changes in existing system programs were required.

Since the new A/A search uses the connection table obtained from blowing up a structure encoded in MCC (mechanical chemical code), the query preprocessor had to be modified to translate query structures into this CT format. Further modification of the preprocessor and search phase will be required to allow the use of D for deuterium, T for tritium and the user-defined element symbols E1, E2, ..., but this effort is being deferred while the current system is documented.

The files which store the structural formula in MCC use a compacted structural formula image (SFI). The search system had to be modified to expand the SFI prior to output. Changes to output programs were also made to provide more space for the A/A search programs and for the MCC to connection table expansion programs.

As a result of these modifications, the entire search system can be accommodated in three phases rather than the previous four.

Query Logical Expression Expansion Program - The program which expands the keys logical expression (KEYS =) and the structure logical expression (STRUCTURE =) in a query has been rewritten and debugged, and documentation is now in progress. This program enables the querist to write more compact logical expressions. Suppose, for example, it is required to retrieve compounds assigned keys K1 and K2 and also compounds assigned keys K1 and K3. Instead of demanding

K1 AND K2 OR K1 AND K3

The user can simply write

K1 AND (K2 OR K3)

which is automatically expanded internally. A command for having the expanded version of the expression printed out on the teletype is also provided.

Additional Teletype Commands - Changes have been incorporated into the teletype command language that allow the user to specify the type of output desired after the accession list size has been printed for him on the teletype. The following are the possible commands:

- (1) OUTPUT PRINTER sends the total output to the Data Products Line Printer.
- (2) OUTPUT STAT restricts the output to statistics giving the total number of answers.
- (3) OUTPUT REGISTRY send the registry numbers and additional local control numbers to the teletype or to the line printer if @OUTPUT PRINTER is also specified.

e.g., if the user types:

@OUTPUT STAT @OUTPUT PRINTER @START	}	the statistics are sent to both the teletype and printer.
@OUTPUT STAT @START	}	statistics go to teletype only.
@OUTPUT PRINTER @START	}	total record goes to line printer.
@OUTPUT REGISTRY @OUTPUT PRINTER @START	}	registry numbers go to line printer. statistics to both teletype and line printer.
@OUTPUT REGISTRY @START	}	registry numbers and statistics go to teletype.
@START		total record punched on Dura paper tape by teletype.

The commands which are typed before @START can be typed in any order. However, "OUTPUT REGISTRY" and "OUTPUT STAT" cancel each other, and the last one typed specifies the kind of output.

The use of PRINT within the query can still be used, but @OUTPUT PRINTER allows the user to decide whether or not to send output to the printer at a later point in the processing of the query.

4. Status of CIDS Compound File

The CIDS file of compounds actually consists of two subfiles. One of these contains about 7,000 compounds from the Edgewood Arsenal Toxicological Information Center (TOXINFO) and about 26,300 compounds from the Chemical-Biological Coordination Center (CBCC). The other subfile is a growing one, known as the Task 07 file, which is under construction at Edgewood Arsenal. Both subfiles will be searchable on the basis of compound structure via the revised (CIDS No. 6) chemical search screens. The Task 07 compounds will be searchable also in terms of the 59 nonstructural descriptors assembled and tagged to the compounds by Edgewood Arsenal.

File building during this report period has been confined to the Task 07 file. As of 30 January 1970, Edgewood Arsenal has transmitted 13,985 candidate compounds for the file and 10,666 of these have been registered. The remainder consists of compounds which were (a) received subsequent to the last registration, or (b) rejected by the screening programs and are in stages of re-typing, or (c) of types currently inadmissible to the system. Analysis of the results of assigning the revised CIDS chemical screens to the 10,666 registered compounds discloses that the file is acquiring the chemical characteristics to be expected of a large unbiased file and is reported on further in Section 6, page 17. This same file received a total assignment of 43,035 Edgewood Arsenal nonstructural descriptors thus producing an average of approximately four nonstructural assignments per compound. Again as expected, the distribution of the total assignments among the 59 nonstructural descriptors varies widely and is recorded in the following tabulation.

<u>Descriptor</u>	<u>Code</u>	<u>Number of Assignments</u>
Applications	AP	7333
Activity Coefficient	AC	zero
Analytical Detection	AD	163
Analytical Determination	AN	4232
Boiling Point	BP	1680
Biological Suppressant	BS	2132
Crystalline Form	CF	1070
Chromatographic Methods	CM	191
Cost	CO	465

(continued)

<u>Descriptor</u>	<u>Code</u>	<u>Number of Assignments</u>
Critical Pressure	CP	5
Color	CR	1669
Critical Temperature	CT	164
Dissociation Constants	DC	7
Derivatives	DV	2540
Dyestuff Application	DA	zero*
Entropy	EN	6
Eye Irritant or Lacrimator	IE	zero*
Electron Spin Resonance Spectrum	ES	4
Free Energy	FE	4
Geometric Isomers	GI	37
Heat Capacity	HC	6
Heat of Dilution	HD	zero
Heat of Formation	HF	9
Heat of Solution	HS	2
Heat of Sublimation	HU	20
Heat of Vaporization	HV	1
Hydrates	HY	63
Ionization Constants	IC (pKa, pKb)	156
Incapacitating Dose (Dosage)	ID	40
Industrial Applications	BA	zero*
Infrared Spectrum	IR	432
Kinetics of Hydrolysis	KH	52
LD50 (Dosage)	LD	517
(Med) Minimum Effective Dose (Dosage)	ME	740
Melting Point	MP	4722
Mass Spectrum	MS	91
Nuclear Magnetic Resonance Spectrum	NS	101
Optical Isomers	OI	zero*
Optical Rotation	OR	296

* Descriptor added subsequent to the original list.
Further assignments expected.

(continued)

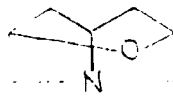
<u>Descriptor</u>	<u>Code</u>	<u>Number of Assignments</u>
Polarography	PO	56
Purification	PU	461
Respiratory Inhibition	RE	7
Refractive Index	RI	958
Raman Spectrum	RS	1
Solvent of Crystallization	SC	2169
Specific Gravity	SG	569
Specific Heat	SH	4
Hammett Sigma Values	SI	28
Solubility	SO	1280
Specifications	SP	148
Surface Tension	ST	9
Suppliers	SU	643
Solvates	SV	7
Synthesis	SY	7121
Synthesis Intermediate or Starting Product	IA	10
Triple Point	TP	zero
Ultraviolet Spectrum	UV	430
Viscosity	VI	109
Vapor Pressure	VP	65

5. Identification of Rings in Cyclic Systems

Considerable thought and effort has been given to the problem of identifying in a cyclic nucleus those rings (or closed paths) which would be of interest to a chemist regardless of whether or not they are the rings prescribed by the Ring Index rules. The recognition of these rings is necessary for the assignment of the CIDS GCN3 (Elementary Ring Population) keys. A solution was sought which would identify the rings of a nucleus by which a chemist would conceivably want to retrieve it, and yet would avoid wherever possible the identification of any other rings which would result in false drops.

The problem turned out to be a very complex one, making a perfect solution unobtainable. However, a highly satisfactory compromise was reached. The solution involves the assignment of GCN3 keys to:

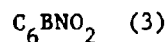
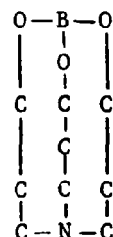
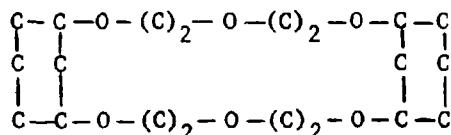
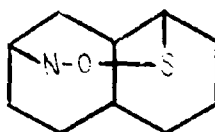
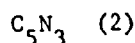
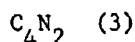
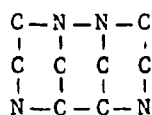
- (1) All rings which are in any smallest set of smallest rings
- (2) All rings of eight atoms or less
- (3) No other rings

The new technique is superior to that described in earlier CIDS reports (CIDS No. 4, p. 64 and CIDS No. 6, p. 18) in that the assignment of elementary ring population (GCN3) keys is independent of system orientation. Thus, for example, the computer automatically assigns one each of the following rings, C_4N , C_5N , and C_7 , to nortropane (RI 1281) and this permits nortropane retrieval either in terms of its Ring Index orientation (C_4N-C_5N) or in terms of its equally valid C_4N-C_7 or C_5N-C_7 orientations. Similarly, no matter how it is drawn, the two-ring system  will have assigned to it, the C_5 , C_6 , and C_7 GCN3 keys and thus will respond to a query posed in terms of C_5-C_6 , C_5-C_7 , or C_6-C_7 .

It requires but a brief excursion into ring systems to appreciate the ring-finding capacity of the technique. Even with such a small system as (RI 10248), for example, the technique finds two C_4N 's, two C_5NO 's, and one each of C_5O , C_4NO , and C_7N . With more complex systems, the ring findings become progressively more abundant thus requiring more storage space in the computer memory and increasing the number of false responses to certain types of queries. It is to avoid these undesirable features that application of the technique is restricted in accord with rules (2) and (3) above.

Regardless of the restrictions, however, it is emphasized that, for any ring system, the technique strives to find as a minimum (a) the set of rings corresponding to the Ring Index (Chemical Abstracts) notation, and (b) any other set of independent rings which has the same numerical ring population (CIDS GCN2 keys) as the Ring Index set.

Additional examples are shown below with the GCN3 rings given:



A program was previously written and tested which finds a "smallest set of smallest rings" in a cyclic compound. Another program, XRINGS, which finds any additional rings to be identified, has now been written, tested, and debugged. A test on a Task 07 file containing 5976 cyclic compounds showed that program XRINGS increased the GCN3 key assignments from 12,611 to 13,181, i.e., by about 5 percent. An algorithm stipulating the basic logic of the program, and the theory behind it, was transmitted to the sponsor prior to writing the program.

The CIDS ring key assignment system, of which XRINGS is a part, can now handle all but a small percentage of the more intricate ring systems; and compounds containing these intricate ring systems will be rejected automatically for intellectual analysis and addition to the file at a future date.

6. New Screen Assignment

All new (CIDS No. 6) screen assignment programs pertaining to organic compounds, including program XRINGS (Section 5 above), have been written and tested, and the screens have been assigned to a file of 7226 Task 07 compounds. These same screens will be assigned to all currently registered (10,666) compounds in the Task 07 file within the next few days and to the total file of about 33,000 TOXINFO and CBCC compounds within the next few weeks.

The computer time required for assignment of keys to the file of 7226 compounds can be broken down into the following averages:

Screens Assigned	Compounds Per Minute	Seconds Per Compound
Ring Keys	57	1.1 sec
Functional Group and Other Keys	8	7.5 sec
Total	7	8.6 sec

An average of 31 keys per compound were assigned, including structural keys, non-structural keys, and registry number keys. This compares with an average of 23 keys per compound for the experimental (CIDS No. 4) structural keys and non-structural keys. The difference is primarily made up of the two registry number keys, the specific cyclic nuclei (SCN) keys (which were specified in CIDS No. 4 but were never actually assigned), additional molecular formula keys, and the assignment of four extra-cyclic carbon (EC) keys to both cyclic and acyclic compounds (their CIDS No. 4 counterparts were assigned only to acyclic compounds).

Three copies of an index listing the keys that were assigned, and the number of assignments for each, were transmitted to the sponsor in order to provide members of Edgewood Arsenal's chemical and programming staffs a better feel for the degree of specificity of each key, and to observe any errors that might be present.

Cursory examination of the index statistics reveals that the file of compounds is acquiring the aspects to be expected of a large unbiased library of compounds. Some examples follow:

(1) Approximately 17 percent of the 7226 file compounds are acyclic, as revealed by the number of assignments (1250) of the A-C=00 key.

(2) The large majority of the cyclic nuclei represent one-ring systems (7393 out of the total of 9435 GCN1 assignments).

(3) The skeleton molecular formula (GCN4) keys were assigned large numbers of times when they involved common ring systems and fewer and fewer times as they involved less and less common systems.

(4) Of the 137 CIDS specific cyclic nuclei (SCN) keys, 102 were encountered, with the greatest number of assignments (5133) going to benzene as expected. Examples of other high scores: pyridine (399), piperidine (293), cyclopropane (285), cyclohexane (293), and indole (135).

(5) 316 of the approximately 550 CIDS specific functional group keys were encountered, with high scores being recorded by common groups such as FG112 (monohalomethyl), FG178 (ether), FG94 (carboxyl), etc., and scores of one only with low frequency groups such as FG3 (arsines), FG71 (nitrolic acids and esters), etc.

(6) Among the saturated monovalent hydrocarbon radicals, HR1 (methyl) was by far the most frequently assigned (5666 times), followed by HR3 (ethyl) (1625 times) and HR10 (propyl) (221 times).

(7) The carbon content of the compounds peaks in the range of C_{10} to C_{14} which is in the same general area reported for large files of compounds.

(8) The molecular formula key assignments are high for those hetero-elements which are expected to occur frequently such as nitrogen, oxygen, the halogens, etc., and low for elements such as manganese, lead, silver, deuterium, etc.

(9) The utility of the nonspecific diatomic functional group (ND) keys is evident from the facts that 20 of the total theoretical number of 66 such keys were assigned and that there was a total of 692 such assignments.

(10) The sparsity of assignment of the nonspecific monatomic (NM) keys is indicative of the comprehensive character of the CIDS lexicon of structural screens.

7. Cathode Ray Tube Research and Development

The program organization of the CRT-terminal system is being altered basically. Occasional past erratic performance was traced to problems in timing and transfer of control, problems that would have been intensified by the addition of hanging bonds into the language. These problems do not in any way affect estimates of feasibility of use of the CRT as a terminal of the CIDS system. They require only more careful application of existing techniques. Progress to date in implementing the altered version is summarized in the following.

The CIDS Graphics Monitor - An operating system has been designed for the DEC-338 programmed buffered display, suitable for the display of chemical compounds and fragments, as used in the CIDS real time search system. The display system consists of a program called the Graphics Monitor which contains subroutines, (such as a routine for displaying a light button or a line of text) and an I/O package to aid in the use of the teletype, dataphone and disk. It also contains features such as automatic light pen tracking and a loader to load program segments and data from the disk. This organization eliminates processing at the level of the primitive display file, i.e., to place an atom on the screen, it is only necessary to call a subroutine in the monitor, specifying the element type and coordinates on the screen. Bonds, light buttons, and even run-standard display files may be entered into the system by the use of such subroutines. As a result, the programs which interact with the user can be fairly short.

Program Segments - The following segments have been written to allow the drawing of a molecule on the screen. They all operate under the Graphic Monitor, and depend on its facilities:

- (a) Display Monitor - permits selection and execution of the functions available to the user through light buttons.
- (b) Drawing Monitor - permits selection of one of the drawing functions, two of which are described in (d) and (e), and the third has not yet been written.
- (c) Draw Atom - drawing of individual atoms on the screen.
- (d) Draw Bond - drawing of bonds on the screen, including closing of rings.
- (e) Atom List - a subroutine for displaying a list of element types, so the user can select the type he wants with the light pen; used in conjunction

- with Draw Atoms, but written as a subroutine so it can be used elsewhere.
- (f) Delete - allows deletion of individual atoms or bonds, or a whole area.
 - (g) Move Area - allows moving a rectangular section of the drawing from one place on the screen to another.
 - (h) Transform - allows reflection of a rectangular section of the drawing about either or both axes.
 - (i) Move Atom - permits moving any individual atom on the screen.

A program has also been written for the processing of queries typed in at the DEC-332 teletype.

Programs to be Written - The program segments for altering the drawing have not been completed. Query processing, text editing and output display also have not been written.

8. Documentation: Query Formulation and Encoding (4)

This document, which is now being processed for distribution as the CIDS No. 7 Status Report, describes and illustrates the rules for encoding queries addressed to the initial model of an operational chemical information and data system (CIDS). The teletype command language for submitting queries to the system was described in an earlier publication and is included in the document as Appendix A. The coding rules constitute an update of the method employed in the earlier experimental system; as a result of extensive experimentation with that system, the capabilities of CIDS have been expanded for the model system, thus necessitating the current revision. In addition to describing the encoding and input of retrieval demands, the document discusses certain features of system operation and use to assist in the formulation of maximally efficient queries.

The chemical features on the basis of which a search of the CIDS file of compounds can be conducted are collectively referred to as the CIDS chemical search components. The complete collection of components, along with the code by which each is represented, is contained in the CID No. 6 report entitled Handbook of CIDS Chemical Search Components. The CIDS No. 7 document discusses the encoding of these components into a valid query and includes the procedure for initiating and modifying file search in response to that query. The CIDS No. 6 report is thus corequisite to CIDS No. 7 for use of the real-time model operational system.

A query coding form has been devised on which the user's question, the pertinent search components, and the fully encoded query are recorded. To simplify reference to the rules during the actual formulation and encoding of a query, the topics in the document have been organized to parallel the coding form. Thus certain sections of the document are of primary interest to the chemist selecting the components appropriate to a given query, while other sections dealing with the encoding of these components are of major concern to computer personnel.

The final section of the document consists of the coding forms for some twenty sample questions exactly as these forms are completed for the operational system. Accompanying each example is an explanation of the strategy employed in the search. Besides providing complete illustrations of query encoding, these examples are designed to suggest to the user a number of techniques for maximizing search efficiency.

9. Documentation: ACT III Chemical Typing Conventions (5)

This document provides detailed instructions for CIDS usage of the DURA chemical typewriter. The document is divided into four sections. Section I deals with preparation of the DURA machine and paper tape. Section II provides general rules and procedures regarding the typing and correcting of a compound record. Section III specifies when, how, and where all nonstructured parts of the chemical record are typed. Section IV provides and illustrates the rules for typing a structured part of the chemical record. In draft form, the document was transmitted to the sponsor under date of 29 December 1969 for review and comment.

10. Documentation: Definition of CIDS Master Files (6)

This document defines the CIDS Master Files at the programmer level and five copies were transmitted to the sponsor under date of 29 December 1969. The documentation includes the format of each data element within the CIDS compound record. Ten important CIDS files are described in terms of the following characteristics:

- (1) sequence of logical records
- (2) variable length or fixed length physical records

(3) logical record description:

- (a) control information present
- (b) data elements present

(4) end-of-reel and end-of-file indicators.

The ten files described are:

- I. CHEMTYPE output, input to ADDMF
- II. Output of ADDMF, input to Ring Screens
- III. Output of Ring Screens
- IV. Input to Registry (potential Registrants)
Registry Master
- VI. Output of Registry (New Registrants)
- VII. Disk Search File
- VIII. Tape Search File
- IX. Key-Address List and INDX (Part of Key Index)
- X. List-of-Addresses File (Part of Key Index)

Sample tape dumps of each file accompany the file descriptions. Included in the documentation are several recommendations for changes in file formats when the data are converted for use on the sponsor's S & E computer.

11. Documentation: Proposed CIDS Chemical Specifications for Polypeptides (7)

This document (a) describes a proposed technique for the molecular and structural representation of polypeptides, (b) identifies the compositional and structural features which must be adequately accommodated for effective automated search and retrieval, and (c) applies the technique to a generous array of sample polypeptides from three different sources. It thus provides the basic information required for the future writing of the necessary computer programs, and was transmitted to the sponsor under date of 21 October 1969.

Because of the cumbersome nature of most polypeptide structures, CIDS finds it advantageous to follow the practice of the international community of biochemists in which the commonly occurring peptide units (amino acid residues) and substituent (usually protective) groups are represented by letter abbreviations rather than by their chemical structures. The commonly occurring peptide

units and substituent groups are listed, along with their abbreviations; and provision is made for the structural representation of any such uncommon units and groups. The general scheme for the structural representation of polypeptides in searchable CIDS format is described and applied specifically to each of the sample polypeptides.

The technique provides for automatic search and retrieval in terms of any one, or any combination, of the following parameters:

- (1) Identification - The generic polypeptide key (PEP), which is assigned to each polypeptide compound, permits (a) confining a search to polypeptides, or (b) including polypeptides in or excluding them from a search.
- (2) Elementary Composition - Molecular formula probing via the existing CIDS Molecular Formula Keys and Statement provisions.
- (3) Polypeptide Size - The total number of peptide units in the polypeptide, whether symbolized or structured.
- (4) The total number of different symbolized peptide units in the polypeptide irrespective of stereo form.
- (5) The identity, but NOT the quantity, of all symbolized peptide units in the polypeptide irrespective of stereo form.
- (6) The quantity (number) of each symbolized peptide unit present in the polypeptide irrespective of stereo form.
- (7) The qualitative presence of one or more structured peptide units in the polypeptide irrespective of stereo form.
- (8) The qualitative presence of any abnormal stereo form of the individual symbolized peptide units. (By an abnormal form is meant a D-, DL-, or allo-form.)
- (9) The qualitative presence of a labeled atom (isotope), including deuterium (D) and tritium (T), anywhere in the compound. Accomplished through use of the existing CIDS MASS Key.
- (10) Structural features of structured peptide units. Accomplished via the existing CIDS Structural Fragment Keys.

Except for elementary composition, search strategy is independent of the symbolized substituent groups. Thus a polypeptide responds to a search regardless of the identity of any substituents it may contain.

Similarly, search strategy is independent of addend compounds thus enabling retrieval of the polypeptide whether registered as such or as a polypeptide-addend compound.

Examination of query responses for a stipulated, partial or total sequence of peptide units is by visual inspection. Subsequent experience may dictate the desirability of computerizing this aspect of search.

12. Documentation: The CIDS Treatment of Synthetic Polymeric Substances (8)

This document describes the recommended technique for the CIDS accommodation of polymeric substances, other than any representative of classes of natural polymers such as polypeptides, polysaccharides, and polynucleotides. In addition to prescribing for molecular and structural representations, the document (1) identifies the compositional and structural features which must be adequately accommodated for effective automated search and retrieval, and (2) illustrates application of the technique to a variety of polymers. The document thus provides the basic information required for the future writing of the necessary computer programs, and was transmitted to the sponsor under date of 28 November 1969.

The document is actually a revised version of an earlier (6/18/69) working paper on the same subject, the primary objectives of the revision being to:

- (1) secure greater compatibility with the practices of the international community of polymer chemists,
- (2) focus on the composition of a polymer, wherever possible, rather than exclusively on the starting chemicals in its manufacture,
- (3) utilize the search advantages afforded by the structural repeating unit wherever one is contained within the polymer,
- (4) adapt the search strategy to probe on any of the composition and structural features in (2) and (3) above.

Provision is made for alternative methods of recording information on molecular formula and structure, and the method of choice for a candidate polymer entry is determined by the kind(s) of information available on it.

Following exhibits of the contents of a CIDS polymer record and the parameters of polymer search, the document concludes with a display of 52 sample polymeric substances in CIDS input format, and the collection is inventoried against a variety of features to serve as a guide to persons responsible for the initial (manual) phase of CIDS input. In arriving at the chemical specifications and at the format for accommodating them at the time of CIDS input, due consideration has been given to expectable user requirements and consultations have been held with computer personnel to assure compatibility with existing CIDS computer programs and with those which have yet to be written.

Currently existing CIDS computer programs are capable of manipulating those polymers which consist of an indefinite number of molecules of a single molecular species, with or without conjuncted H_2O . In other words, the molecular formula of the polymer must be exactly $(monomer)_a$ or $(monomer)_a \cdot H_2O$. Obviously, these represent homopolymers formed either by the opening of unsaturated linkages, e.g., the vinyl and acrylic types, or the rupturing of ring structures, e.g., the ethylene oxide type. Any such polymers admitted to the CIDS file now, however, will have to be deleted and re-entered, in accord with the revised format, when the final computer programs embracing all polymers are available.

13. Documentation: The CIDS Treatment of (1) Metal-Containing Organic Compounds and (2) Esters (9)

This document provides the chemical specifications for the CIDS treatment of all types of metal-containing organic compounds. Since the acid moiety of an ester often contains a metal as the central atom, e.g., ethyl chromate, it was deemed advisable to include an additional section in the document summarizing the CIDS treatment of all esters. The document was completed on 20 January 1970 and copies were transmitted to the sponsor on 2 February.

The total class of metal-containing organic compounds is a broad one which can be conceived as embracing several structurally specific subclasses (types), e.g., normal and acid salts, basic salts, metal derivatives of compounds other than systematic acids, metallocenes, coordination complexes, etc. In the interest of useful discrimination, CIDS takes this into account and (a) subdivides the total classes into types deemed expedient for the system, and (b) utilizes

structuring techniques and search strategies which are appropriate to the different types and are compatible with conventions familiar to the chemical community. Although some of the types are amenable to currently existing (CIDS No. 6) search screens and have been admitted to the growing CIDS file of compounds for quite some time, nevertheless, in the interest of completeness, it was deemed appropriate to describe these also in the document.

Each type of compound is defined and several illustrations of the chemical contents of the CIDS compound record are provided for each. The types which are currently admissible (responsive to CIDS No. 6 search screens) are identified. The chemical specifications for the other types are set forth, but the computer programs necessary for registry and search have yet to be written.

14. CIDS Documentation Scheduling Requirements

Through a series of conferences with the Project Officer, a schedule has been arrived at for the production of various elements of CIDS documentation, some of which have already been completed and forwarded to the sponsor. Priority is currently being given to the production, by 15 March 1970, of interim documentation of the overall system in detail sufficient to disclose to chemically knowledgeable and computer experienced personnel the types and magnitude of effort required to complete the system and convert it to a computer to be designated by the sponsor. By completing the system is meant writing and incorporating the computer programs necessary to accommodate those types of compounds which are currently inadmissible to the CIDS file. Prominent among these types are: polypeptides (Section 11 above), polymers (Section 12 above), metal-containing organics (Section 13 above), and inorganic compounds (chemical specifications ready for documentation).

The immediate emphasis on the interim documentation requires that the completion dates originally envisaged for some of the contract tasks be deferred, and the sponsor has been advised of the new dates.

LITERATURE CITED

1. Clarence T. Van Meter, Eric N. Goldschmidt, Margaret Milne, Handbook of CIDS Chemical Search Components, CIDS No. 6, University of Pennsylvania, Philadelphia, Pa., December 1968
2. Clarence T. Van Meter, David Lefkovitz, Ruth V. Powers, An Experimental Chemical Information and Data System, CIDS No. 4, University of Pennsylvania, Philadelphia, Pa., January 1967
3. Morris Plotkin, Ring Finding Algorithm (memorandum report), University of Pennsylvania, Philadelphia, Pa., October 1969
4. Margaret Milne, Paul R. Weinberg, Query Formulation and Encoding, CIDS No. 7, University of Pennsylvania, Philadelphia, Pa., November 1969
5. Nancy Hanp, Helen Hill, R. T. Swem, ACT III (DURA) Chemical Typing Conventions (draft), University of Pennsylvania, Philadelphia, Pa., December 1969
6. Ruth V. Powers, Definition of CIDS Master Files, University of Pennsylvania, Philadelphia, Pa., December 1969
7. Proposed CIDS Chemical Specifications for Polypeptides, Project CIDS, University of Pennsylvania, Philadelphia, Pa., October 1969
8. The CIDS Treatment of Synthetic Polymeric Substances, Project CIDS, University of Pennsylvania, Philadelphia, Pa., November 1969
9. The CIDS Treatment of (1) Metal-Containing Organic Compounds (2) Esters, Project CIDS, University of Pennsylvania, Philadelphia, Pa., January 1970

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) UNIVERSITY OF PENNSYLVANIA Philadelphia, Pennsylvania 19104		2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	
		2b. GROUP NA	
3. REPORT TITLE A CHEMICAL INFORMATION AND DATA SYSTEM			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Semiannual Report, 1 September 1969 - 31 January 1970			
5. AUTHOR(S) (First name, middle initial, last name) Van Meter, Clarence T., Powers, Ruth V., Hill, Helen N., Milne, Margaret, Hamp, Nancy, Chen, T. C., and Plotkin, Morris.			
6. REPORT DATE 31 January 1970	7a. TOTAL NO. OF PAGES 34	7b. NO. OF REFS 9	
8a. CONTRACT OR GRANT NO. DAAA15-69-C-0140	8b. ORIGINATOR'S REPORT NUMBER(S)		
8c. PROJECT NO. c. Task: 2P062101A72702	8d. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)		
10. DISTRIBUTION STATEMENT This document is subject to special export controls and each transmittal to a foreign government or a foreign national may be made only with prior approval of the Commanding Officer, Edgewood Arsenal, ATTN: SMUEA-TSTI-T, Edgewood Arsenal, Maryland 21010			
11. SUPPLEMENTARY NOTES Chemical information data systems		12. SPONSORING MILITARY ACTIVITY Edgewood Arsenal Technical Support Directorate, Edgewood Arsenal, Maryland 21010 (Stanley Goldberg, Proj. O., Ext. 6126)	
13. ABSTRACT This document describes research activities in progress under Project CIDS at the University of Pennsylvania through which a model operational chemical information and data system has been developed. The document discloses the present state of CIDS file building and reports the successful incorporation of the revised chemical search screens into the system and the establishment of chemical specifications for handling certain currently inadmissible classes of compounds. Various improvements in search and file building techniques are reported, including an acceptable method for analyzing complex ring systems, and the status of R&D with the cathode ray tube as an input-output device is described. Also included are reports of recent CIDS computer documentation and a description of the formal CIDS No. 7 report on query formulation and encoding.			
14. KEYWORDS:			
Data system		Encoding	
Chemical information		Connection table compaction	
CIDS chemical search keys		Ring-finding algorithm	
Cathode ray tube		Master file documentation	
Polypeptides		Query formulation	
Synthetic polymers		Compound files	
Organo-metallic compounds		Chemical typing conventions	
Atom-by-atom search		CIDS interim documentation	
Nonstructural descriptors		Expansion of logical expressions	
Teletype command language		Mechanical Chemical Code	

DD FORM 1473

REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR USE.

UNCLASSIFIED

Security Classification